Appendix H
2009 ISTEP+ Reliability and Validity Report

This report describes some of the evidence that establishes the degree to which the ISTEP+ tests are reliable and valid. These tests were designed to measure students' skills in the domains of English/Language Arts, Mathematics, and Science as defined by the *Indiana Academic Standards.*

Reliability

Test scores always contain some amount of measurement error. This kind of error can be random or systematic. Standardization of assessments is meant to minimize random error that occurs because of random factors that affect a student's performance on the test. Systematic errors are inherent to examinees and are typically specific to some subgroup characteristic (i.e., students who need accommodations but are not offered them). Reliability refers to the degree to which students' scores are free from such effects and provides a measure of consistency. In other words, reliability helps to describe how consistent students' performances would be if given the assessment over multiple occasions.

For the ISTEP+, several measures of reliability are available. First, the tests are administered in standard fashion to all students. When students need accommodations, such accommodations are provided with specific guidance from the ISTEP+ Program Manual (www.doe.in.gov/achievement/assessment) that describes details about the tests, as well as specific administration policies, procedures, and accommodation guidelines.

Item-Level Reliability

Item-specific reliability statistics include inter-rater reliability, point biserial or item-test correlations, and differential item functioning (DIF) or item bias. The inter-rater reliabilities of CR items rely heavily on the solid and consistent training of the Handscorers, as was described in Section 4 – Scoring. Statistical data are presented in terms of the kappa and intraclass correlations as ways to measure the consistency (reliability) of the scores. Tables 8−11 provide the relevant inter-rater reliability statistics. In general, the values are within acceptable limits. The lowest statistics fall on one SS field test item that presents intraclass statistics of 0.69 and kappa statistics of 0.37. Intraclass correlations for all items range from 0.74 to 0.97 with a mean of 0.89 (ELA); from 0.79 to 1.00 with a mean of 0.94 (MA); from 0.86 to 0.98 with a mean of 0.94 (SC); and from 0.69 to 0.95 with a mean of 0.87 (SS). Kappa statistics range from 0.47 to 0.93 with a mean of 0.78 (ELA); from 0.58 to 1.00 with a mean of 0.87 (MA); from 0.72 to 0.96 with a mean of 0.87 (SC); and from 0.37 to 0.90 with a mean of 0.73 (SS). These values are within acceptable limits.

The point biserial or item-test correlation, a type of internal consistency measure, is one measure of the correlation between each item and the overall test as was described in Section 6–Methods, results of which were described in Section 7-Results. The item-test correlations for each content area, grade, and item type are shown in Table 18. The correlations for operational items range from 0.14 to 0.58 (ELA); from 0.13 to 0.70 (MA); from 0.11 to 0.53 (SC); and from 0.15 to 0.53 (SS). The correlations for field test items range from 0.08 to 0.58 (ELA); from 0.01 to 0.66 (MA); from 0.03 to 0.58 (SC); and from 0.05 to 0.52 (SS). Field test items show much lower ranges, and some field test items that had negative correlations were removed from the pool of items. All items with item-test correlations lower than 0.30 have been reviewed by Research, Publishing, and the IDOE and none of the items were mis-keyed or had possible multiple correct answers, as might be indicative of such low correlations. Certainly, any items with extremely low point biserials that may remain in the item pool will be avoided on future operational forms.

DIF statistics (described in Section 6–Methods and Section 7–Results) provide a measure of the systematic errors by subgroups that are specifically attributed to some bias or systematic over- or under-representation of subgroup performance when compared to total group performance. As mentioned and apparent in Tables 23 and 24 (last rows), only about 7% of the operational items exhibited gender or ethnic DIF at the moderate and large levels; and for field test items, only about 6% exhibited moderate or large levels of gender or ethnic DIF.

Test-Level Reliability

Total test reliability statistics (alpha and SEMs) measure the level of consistency (reliability) of performance over all test questions in a given form, the results of which imply how well the questions measure the content domain and could continue to do so over repeated administrations. Total test reliability coefficients (in this case measured by Cronbach's alpha (α, 1951), may range from 0.00 to 1.00, where 1.00 refers to a perfectly reliable test. The ISTEP+ reliability data are based on Indiana-specific representative samples from each grade (the scaling sample), and the results for 2009 are typical of the results obtained for all previous ISTEP+ operational tests. The total test reliabilities of the operational forms were evaluated first by Cronbach's α (Cronbach, 1951) index of internal consistency. The specific calculation for Cronbach's α is calculated as

$$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum \hat{\sigma}_i^2}{\hat{\sigma}_X^2}\right) \quad (8.1)$$

where k is the number of items on the test form, and $\hat{\sigma}_i^2$ is the variance of item i and $\hat{\sigma}_X^2$ is the total test variance. Achievement tests are typically considered of sound reliability when their reliability coefficients are in the range of 0.80 and above.

Table 42 shows the reliability coefficients for each scored test form, containing only operational items, for each grade and content area for both Fall 2008 (and from 2007 for grade 8 content that was tested in grade 9 of 2007) and Spring 2009. Alpha reliability coefficients for Spring are quite similar to Fall, and ranged between 0.87 (grade 5 SS) and 0.93 (grade 8 MA). Such a range is indicative of the high reliability of ISTEP+ tests. As is evident in Tables 29−32, for Spring 2009 state and subgroup data, the coefficients are quite high and similar to the state even at the subgroup levels. Specifically, the average (and range) of the state level reliability coefficients for each content area are as follows: ELA 0.91 (range 0.88−0.94), MA 0.91 (range 0.88−0.95), SC 0.87 (range 0.81−0.92), and SS 0.86 (range 0.79−0.91). At the subgroup level, the lowest reliabilities (0.79 and 0.81) were found for the LEP students in grade 5 SS and grade 6 SC, respectively.

The SEM is another measure of reliability and is a direct estimate of the degree of measurement error in students' total scores at the total test level (per the alpha reliability coefficient) and at the total or scale score level. The SEM represents the number of score points about which a given score can vary, similar to the standard deviation of a score; the smaller the SEM, the smaller the variability of the estimate, and the higher the reliability. The total SEMs are computed with the following formula:

$$SEM = SD\_TT(\sqrt{1-\hat{\alpha}}) \quad (8.2)$$

The SEMs for each scale score are computed with the following formula:

$$SEM = SD\_SS(\sqrt{1-\hat{\alpha}}) \quad (8.3)$$

where SD_TT is the standard deviation for the total test and SD_SS is the standard deviation of the scale score; $\hat{\alpha}$ is the result of the calculation of Cronbach's α above. The total test SEMs for each test form are provided for each grade and content at the state and subgroup levels in Tables 29–32. Scale score specific SEMs are given in Tables 43–46, which also provide the raw scores associated with each scale score. Please note that ISTEP+ uses pattern scoring and does **not** use raw score-to-scale score tables; the raw scores in the tables should therefore be interpreted with pattern scoring in mind.

Proficiency-Level Reliability

One of the cornerstones of the NCLB Act (2002) is the measurement of Adequate Yearly Progress (AYP) for states with respect to the percentage of students at or above the academic performance standards established by states. Because of a heavy emphasis on moving all students to or above the "Proficient" category by year 2014, the consistency and accuracy of the classification of students into these performance categories is of particular interest.

The statistical quality of cut scores that define the proficiency levels in which students are placed per their performance serves as additional validity evidence. Details about the Cut Score Setting Workshop and Bookmark procedure used to set the cut scores are given in the *ISTEP+ Cut Score Setting Technical Report* (CTB, 2009). It may be useful to note here that the Bookmark procedure (Mitzel, Lewis, Patz, & Green, 2001) is a well-documented and highly regarded procedure that has been demonstrated by independent research to produce reasonable cut scores on tests across the country.

It is also important to review the specific scale score SEM for each cut score. Table 47 shows the SEMs estimated for each of the Spring 2009 cut scores for each content area and grade. Comparison of these SEMs to the SEMs associated with other ISTEP+ scale scores for each test (shown in Tables 43–46) reveal that these values are almost always among the lowest, meaning that the ISTEP+ tests tend to measure most accurately near the cut score. This is a desirable quality when cut scores are used to classify examinees. (Note that every scale score possible, sometimes including the cut score, is not shown in Tables 43–46; there are more scale scores possible at each raw score than can be shown in these tables.)

Not only is it important that the amount of measurement error around the cut score be minimal; also important is the expected consistency with which students would be classified into performance levels if given the test over repeat occasions.

Classification consistency is defined as the extent to which two classifications of a single student agree from two independent administrations of the same test (or two parallel forms of the test). Classification consistency and accuracy are additional measures of reliability as well as validity. Reliability coefficients, such as Cronbach's alpha, are used to check for the internal consistency within a single test. Test-retest reliability requires two administrations of the same test which requires another test as an external reference. When retesting students is not feasible, classification consistency is a viable and often utilized alternative. Consistency in the classification sense represents how well two forms of an assessment with equal difficulty agree (Livingston & Lewis, 1995). It is estimated using actual response data and total test reliability from an administered form of an assessment, from which two parallel forms of the assessment are statistically modeled and classifications compared.

Classification accuracy is defined as the agreement between the actual classifications using observed cut scores and true classifications based on known true cut scores (Livingston & Lewis, 1995). It is common

to estimate classification accuracy by utilizing a psychometric model to find true scores corresponding to observed scores.

In other words, classification *consistency* refers to the agreement between two observed scores, while classification *accuracy* refers to the agreement between the observed score and the true score. A straightforward approach to classification consistency estimation can be expressed in terms of a contingency table representing the probability of a particular classification outcome under specific scenarios. For example, below is a contingency table of (H+1) x (H+1), where H is the number of cut scores, such that two cut scores yield a 3x3 contingency table.

	Level 1	**Level 2**	**Level 3**	**Sum**
Level 1	P_{11}	P_{21}	P_{31}	$P_{\cdot 1}$
Level 2	P_{12}	P_{22}	P_{32}	$P_{\cdot 2}$
Level 3	P_{13}	P_{23}	P_{33}	$P_{\cdot 3}$
Sum	$P_{1\cdot}$	$P_{2\cdot}$	$P_{3\cdot}$	1.0

To report classification consistency, Swaminathan, Hambleton, and Algina (1974) suggest using Cohen's kappa (1960):

$$\text{kappa} = \frac{P - P_c}{1 - P_c}, \qquad (8.4)$$

where *P* is defined as sum of diagonal values of the contingency table (shaded above) and P_c is the chance probability of a consistent classification under two completely random assignments. This probability, P_c, is the sum of the probabilities obtained by multiplying the marginal probability of the first administration and the corresponding marginal probability of the second administration:

$$P_c = (P_{1\cdot} \times P_{\cdot 1}) + (P_{2\cdot} \times P_{\cdot 2}) + (P_{3\cdot} \times P_{\cdot 3}) \qquad (8.5)$$

Kolen and Kim (2005) suggested a method for estimating consistency and accuracy that involves the generation of item responses using item parameters based on the IRT model (see also Kim, Choi, Um, & Kim, 2006, as well as Kim, Kim , & Barton, 2007). Two sets of item responses are generated using a set of item parameters and an examinee's ability distribution from a single test administration. These two sets of item responses are considered as an examinee's responses on two administrations of the same form. The procedure is described below and is implemented with the KKCLASS software (Kim, 2005).

Step 1: Obtain item parameters (**I**) and ability distribution weight ($\hat{g}(\theta)$) at each quadrature point from a single test.

Step 2: Compute two scale scores at each quadrature point. At a given quadrature point θ_i, generate two sets of item responses using the item parameters from a test form, assuming that the same test form was administered twice to an examinee with the true ability θ_i.

Step 3: Construct a classification matrix at each quadrature point. Determine the joint event for the cells (as illustrated in the table above) using the raw scores obtained from Step 2.

Step 4: Repeat Steps 2 and 3 *R* times and get average values over *R* replications.

Step 5: Multiply distribution weight ($\hat{g}(\theta)$) by average values in Step 4 for each quadrature point, and sum across all quadrature points. From this final contingency table, classification consistency indices, such as consistency agreement and kappa, can be computed.

Step 6. Because examinees' abilities are estimated at each quadrature point, this quadrature point can be considered the true score. Therefore, classification accuracy is computed using both examinees' estimated abilities (observed scores) and quadrature point (true score).

Table 48 shows classification consistency and classification accuracy indices. Note that the values of all indices depend on several factors, such as the reliability of the actual test form, the distribution of scores, the number of cut scores, and the location of each cut score. The probability of a correct classification (Consistency) is the probability that the classification the student received is consistent with the classification that the student would have received on a parallel form; in other words, that the classification is correct. This is akin to the exact agreement rate in inter-rater reliability, and the expectation is that this probability would be high. The average Consistency is 0.88 across all grades and content areas, and ranges from 0.77 (SC grade 6 and SS grade 7 across both cut scores) to 0.98 (MA grade 8, Pass Plus cut score).

The probability of a correct classification by chance (Chance) is probability that the classification is correct and is due to chance alone. The probability of Chance is estimated under a complete random assignment procedure using the marginal distribution of each form. The Chance probabilities are expected to be low, and in this case are lowest where the Consistency is highest. Average Chance values across all grades and content areas is 0.60 and ranges from 0.35 (SC grade 7, all cuts) to 0.91 (MA grade 8, Pass Plus cut score).

Cohen's kappa (Kappa) provides the same type of reliability or agreement statistic as described previously in discussing inter-rater reliabilities. In this context, it represents the agreement of the classifications between the two parallel forms with the consideration of the probability of a correct classification by chance (Consistency - Chance)/(1 - Chance). In general, the value of Kappa is lower than the value of Consistency because the probability of a correct classification by chance is greater than 0. This is true of the ISTEP+ data in Table 48. Average Kappa is 0.70 and ranges from 0.59 (MA grade 7, Pass cut score) to 0.81 (ELA grade 8, Pass cut score) over all grades and content areas.
Consistency and accuracy are important to consider together. The probability of accuracy (Accuracy) represents the agreement between the observed classification, based on the actual test form, and true classification given the modeled form. The average Accuracy is 0.88, ranging from 0.64 (MA grade 7, across both cut scores) to 0.98 (MA grade 8, Pass Plus cut score). Finally, Table 48 provides the probability of false positives (FP) and false negatives (FN) as measures of error in the data table, and these are low (no greater than 0.07 and 0.36, respectively), as expected.

Classification consistency and accuracy matrices are also provided (see Table 49). These provide probabilities of classification across observed and expected classification. The diagonals represent probabilities for the classification or accuracy when both the observed and expected classifications were the same, and when the off-diagonals were off by one or two proficiency levels. In almost every case, the diagonal probabilities are higher than the off-diagonals, which is consistent with the Consistency and Accuracy data provided.

Validity

Validity refers to the degree to which theory and evidence indicate that test scores support the meaning and use of the scores as intended (AERA, APA, and NCME, 1999). Basically, "validity is the ongoing trust in the accuracy of the test, the administration, and interpretations and use of results" (Barton, 2008). Test validation is therefore an ongoing process of gathering evidence from many sources to evaluate the trustworthiness of the desired score interpretation or use. This evidence is acquired from studies about

the content of the test, how the test was developed, the blueprints, the alignment, and so forth, to how the procedures and processes support the trust in the data integrity, quality of scoring, psychometric analyses, and reporting. Additionally, reliability is a necessary element for validity. Inferences from test scores cannot be valid if they are not also reliable.

Exploratory and Confirmatory Factor Analyses

Exploratory and Confirmatory Factor Analyses (CFA) were conducted to investigate potential evidence to further support the validity of the ISTEP+ test scores for the total population, and then by SPED, LEP, and accommodated subgroups. The subgroups were chosen such that the students within each group may have characteristics that could contribute to issues of access and/or for whom the test measures construct irrelevant variances. A variety of criteria are used conjunctively to evaluate the assumption that each test for each grade and content area measures a single (unidimensional) construct (e.g., MA, ELA, SC, or SS). In factor analyses, the "construct" is referred to as a factor. The analyses help to organize the data such that relationships defined as factors are illuminated. If the data are essentially unidimensional, a single factor should account for most of the variation in the data.

Accordingly, a unidimensional factor model was tested using polychoric correlation coefficients against the obtained covariance matrix[1] using maximum likelihood estimation (Bentler & Bonett, 1980, Jöreskog, & Sorbom, 1989) for each grade and content area for the total population and each subgroup using SAS version 9.1. The polychoric correlation is most appropriate when variables are dichotomous or ordinal and together are assumed to reflect a single underlying construct (Byrne, 1998).

First, the factorability of the correlation matrix was examined before conducting the CFA (i.e., Is the data adequately correlated and thus analyzable or "factorable" to move forward?). The Kaiser-Meyer Olkin (KMO; Kaiser, 1970, 1974) measure of sampling adequacy was used through an Exploratory Factory Analysis (EFA) procedure to evaluate the strength of the linear relationship among the items within each correlation matrix. KMO values in the 0.90 and greater range are considered "marvelous" according to Kaiser's (1974) criteria. As shown in Tables 50 and 51, KMO values for the total group ranged from 0.96 to 0.98, and, for each subgroup: from 0.94 to 0.97 (SPED), from 0.90 to 0.96 (LEP), and from 0.92 to 0.96 (Accommodated). That all the KMO values are in the "marvelous" range suggests that the matrix is appropriate for CFA for each analysis.

As a rough estimate of the number of factors (dimensions or constructs) that might be present in the data, the Kaiser criterion of computing the eigenvalues for the correlation matrix was examined next. Eigenvalues represent how much variability is accounted for by each factor not in sum, but out of the total amount of variance, which means there will be times the percentages can be greater than 100%. Tables 50 and 51 also show the total amount of variance that exists in each form, as well as the percent of variance accounted for by the initial eigenvalue. For the total group analyses, the first eigenvalue's measure of the amount of variance in relation to the total variance is 87−96% (ELA), 74−89% (MA), 99−104% (SC), and 101−105% (SS). The range of variance by the first eigenvalue in each content area and subgroup is as follows: SPED: ELA 84–90%, MA 72–84%, SC 97–101%, and SS 99–102%; LEP: ELA 77–85%, MA 67–82%, SC 90–92%, and SS 90–94%; Accommodated: ELA 77–89%, MA 68–81%, SC 97–102%, and SS 97–100%. Such values indicate one major factor is present in each of the content

[1] The variance-covariance matrix, as opposed to the correlation matrix, is most appropriate for CFA (Cudeck, 1989).

assessments. It is interesting to note that the MA range of variance is slightly lower than the other content areas for the total population and each subgroup.

As a rule, "essential unidimensionality" is assumed when the ratio of the first eigenvalue to the second eigenvalue is at least three. The final column of Tables 50 and 51 provides the ratio of the first and second eigenvalues. All grades and content areas for the total population and each subgroup have no ratios less than three; therefore, the ISTEP+ tests are demonstrating essential unidimensionality per the eigenvalue ratio criterion.

An additional available criterion used in EFA to judge the number of factors present is the scree test (Cattell, 1966) of eigenvalues plotted against factors. Examinations of the scree plots for all grades and content areas for the total population and each subgroup indicated a single factor model is present and similar patterns between the total population and subgroups.

Next, the CFAs were run on each test form for each group. In the CFA, a collection of goodness-of-fit indices are used to assess the fit of a unidimensional factor model to the observed data. In other words, does a model that imposes a single factor (from the EFA results) bear out in the observed data through a confirmation or CFA? The indices and relevant criteria reviewed include:
(a) the root mean square of approximation (RMSEA; Steiger & Lind, 1980), where RMSEA values below 0.10 indicate a "good fit" to the data and values below 0.05 indicate a "very good fit" to the data (Steiger, 1990);
(b) the comparative fit index (CFI; Bentler, 1990);
(c) the non-normed fit index (NFI; Bentler and Bonett, 1980), also referred to as the Tucker-Lewis index, where larger CFI and NFI values (i.e., values above 0.90) are interpreted as indicating a "good fit" to the data; and
(d) the chi-square test (χ^2) of fit between the predicted and obtained covariance matrices such that a nonsignificant chi-squared value (χ^2) is the criterion.

While chi-square statistics are traditionally presented in such analyses, it is well known that chi-squared values are often erroneously significant with large samples, such as in the case of ISTEP+ data. Therefore, caution should be taken when used for assessing model-data fit for these data; presentation of the information is typical.

Tables 52 and 53 provide the specific values for each index described. In summary, the RMSEA values for each grade and content area and across all groups are all below 0.04 and therefore considered a "very good fit." CFI and NFI values fall in the following ranges:
Total Group: 0.86−0.92 (ELA), 0.75–0.86 (MA), 0.91–0.95 (SC); and 0.92–0.95 (SS)
SPED: 0.84–0.89 (ELA), 0.70−0.83 (MA), 0.90–0.94 (SC); and 0.89–0.93 (SS)
LEP: 0.78−0.90 (ELA), 0.68−0.84 (MA), 0.81–0.92 (SC); and 0.82–0.92 (SS)
Accommodated: 0.79−0.89 (ELA), 0.67−0.80 (MA), 0.88–0.93 (SC); and 0.86–0.91 (SS)

The CFI and NFI values for all content areas dip below 0.90 in most cases and groups, except for the total group in SC and SS. Each chi-square is showing significance ($p<0.001$); however, it is highly likely that the very large sample sizes are contributing to the significance.

Summary inspection across all the criteria - variance, ratio of eigenvalues, scree plots, and goodness-of-fit indices - seems to indicate that the tests for each grade and content area, and for each subgroup, are essentially unidimensional. It will be important to review the relationships of factors particularly in MA in

conjunction with all other data, particularly where items may be dependent (for example, where all CRs are scored twice).

In order to support the valid interpretations and uses of the results, the teachers are provided access to student responses for all open-ended items administered in the first test window (http://www.doe.in.gov/achievement/assessment), and *IReleased Items and Scoring Notes* (same website) for each grade and content area, which provide a brief descriptions of the types of questions assessed by each content area, short summaries of scoring rules utilized by the Handscorers, access to the rubrics used to score student responses, copies of the released open-ended items, and anchor papers used by the Handscorers to distinguish between papers with different scores. Teachers are also provided a *Guide to Test Interpretation* for all grades and content areas (http://www.doe.in.gov/achievement/assessment). The *Guide to Test Interpretation* contains helpful tips on the types of scores and data reported, a brief description of such concepts as IRT and pattern scoring, and guidance on how to interpret various scores and aggregations of scores at various levels.